Digital Forensics: Project #3
Face detection

Due on July 2020

Prof. Simone Milani

Elena Camuffo 1234370

Contents

1 Introduction

The face detection and recognition strategy is a widely used and spread technology in continuous development, which is employed in a huge number of different applications, from mobile devices' cameras to the most advanced video sourveillance systems.
The face detection task was carried out for years and can make use of different techniques, from the geometric based methods in the early 90s, to most sofisiticate algorithms among which we can identify *Viola and Jones algorithm* in the 2000s, to the recent advanced deep learning techinques of these days.

According to the literature the passages needed to perform well the face recognition task are five and they are listed in the following [1]:

1. **Face Localization.** The region where the faces are located ae identified and cropped from the remaining part of the image. The task is carried out by algorithms like *Viola and Jones*.
2. **Alignment and Normalization.** The faces are normalized from a geometrical p.o.v. to canonical coordinates, usually by means of facial landmarks.
3. **Face Processing.** The facial pose and illumination are here compensated and the *data augmentation* trick can be added here.
4. **Feature Extraction.** Features are extracted from each face in order to get a unique representation of each sigle person.
5. **Face Comparision.** Features are compared in order to understand who is the person in the picture, selecting among the ones prensent in the dataset.

These passages can be in many cases merged and overlapped, depending on the structure of the built system.

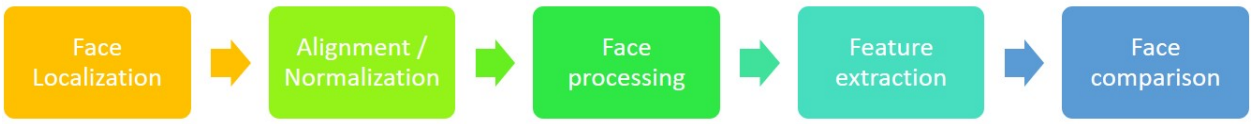


Figure 1: The five passages of face recognition

1.1 Project Overview

In this paper we are going to analyze step by step a face recognition system based on a dataset composed of more than 30 thousands of images representing 13 different people from the Hollywood world, i.e. *Adam Sandler*, *Alyssa Milano*, *Bruce Willis*, *Denise Richards*, *George Clooney*, *Gwyneth Paltrow*, *Hugh Jackman*, *Jason Statham*, *Jennifer Love Hewitt*, *Lindasy Lohan*, *Mark Ruffalo*, *Robert Downey Jr.* and *Will Smith*.

The structure of the system follows roughly the five main passages with some variations and additions.

- The people faces are firstly pre-processed achieving a set of normalized 128x128 greyscale images (section 2).
- Then a convolutional neural network is built (section 3) and it is trained with part of the images.
- The testing is performed on the remaining set of images and also on some modified versions of those images where the faces are partially hidden (half faces).

- Finally also some videos of the people involved are used to test the model in a more challenging way, extracting and processing in real time the faces present in each frame of the video.

To have an evaluation of the model at intermediate levels some measures are used. In particular the accuracy measure, computed as:

$$accuracy = \frac{t_p}{N}$$

where t_p is the number of well classified samples and N is the total number of samples.
The accuracy represents a parameter on which we can base to estimate the reliability of the model, computing it on the training set (and validation split) during the training process, and on the test sets when the model is just built.

The software used are **Matlab** (the provided code of the *prepareDataset* function) for the initial operations and for building half faces, and **Python** for all the other tasks.

2 Pre-processing

The data pre-processing is the first step that has been carried out and consists in the computation and manipulation of the original images in order to obtain a set of standard 128x128 greyscale images.

2.1 Dataset extraction

The images are loaded with the **Matlab** function *prepareDataset* and split into *training* and *test* sets. The code provided is here a little modified to build in this step also the modified datasets, discussed in the following paragraph, making use of the images belonging to the training test sets, keeping them separate.
The dataset is augmented of a factor 10 applying random geometric transforms to the images, in order to enhance the quantity of the images available to feed the cnn (section 3).

(a) Original (b) Masked mouth (c) Masked (d) Blurred eyes (e) Darken eyes

Figure 2: The faces computed in the image pre-processing

2.2 Dataset masking

The modified test sets build versions of the images with partially covered faces adding some elements on it that simulate in someway something that the people can wear ordinarily, i.e.:

1. *Glasses simulation* blurring the region where the eyes are (fig. 2d).
2. *Sunglasses simulation* adding two dark circles ovelayed on the eyes (fig. 2e).

3. *Mask simulation* adding a simple rectangle on the mouth (fig. 2b).

4. *Coronavirus Medical Mask simulation* overlaying an image of a medical mask on the mouth and the nose, as for coronavirus (fig. 2c).

2.3 Dataset normalization

Then the sets are furtherly processed with **Python**, making use also of various tools like *OpenCV* and *Pillow* for the image processing. The images of the training and test sets are here flattened to greyscale and resized to 128x128.

3 Training

The training is performed using a Convolutional Neural Network (CNN) built in **Python** from scratch with the *Keras framework* inside *Tensorflow*[1].

3.1 CNN Model

The model is shown in figure 3. It is composed of 2 convolutional layers each followed by a max-pooling one, with filters of size 5x5; they are furtherly followed by a flattening and 2 feedforward layers at the output perform the classification.

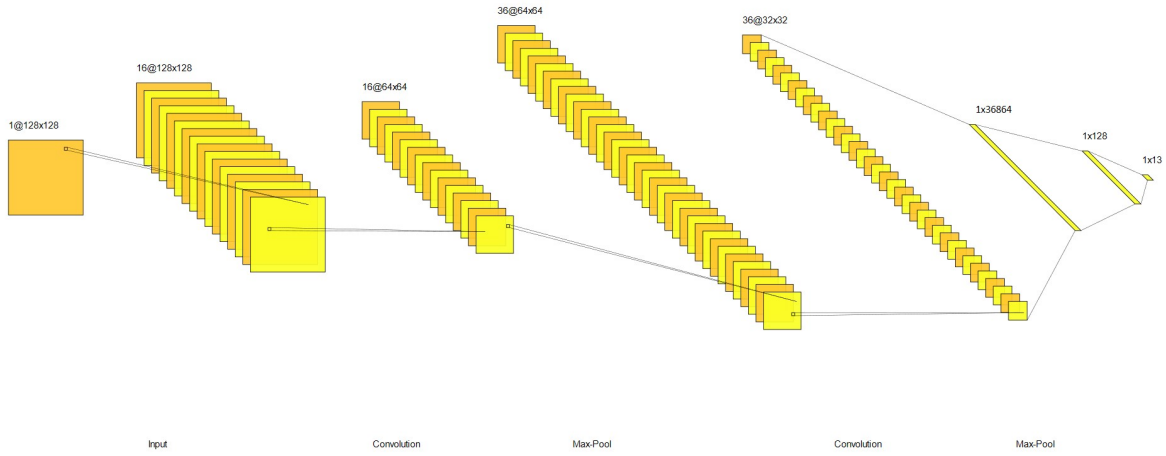


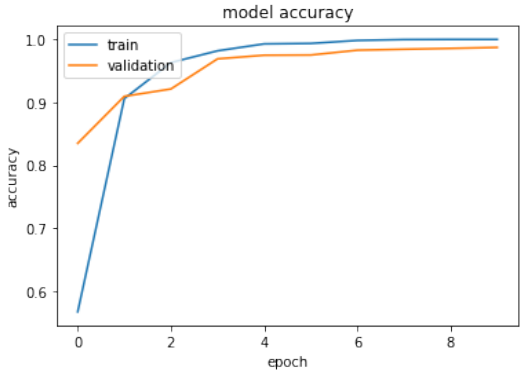
Figure 3: The structure of the convolutional neural network built for the task.

3.2 CNN Classification

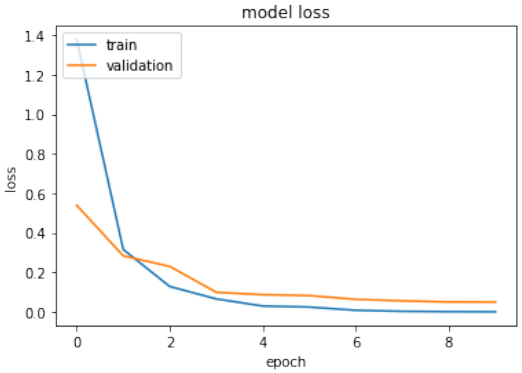
The CNN is feed with the training set and it is trained for *10 epochs* with a learning rate of $\eta = 0.001$ and batch size of 128 samples. In addition, the training set is split into training and *validation* sets (validation split of 20%) to achieve better results. As the learning curve reported in figure 5a shows, the accuracy increases up to 100% for the training samples and almost 98% for the validation.

[1]The model is a modified version of the one of the Keras Tutorial of the Computer Vision course [4]

As regards the loss, it is computed with a *sparse categorical cross entropy* function and figure 5b shows that its behaviour is decreasing over time, down to zero.
Finally *adam* optimizer is used to speed up the convergence.



(a) Accuracy achieved during the training.

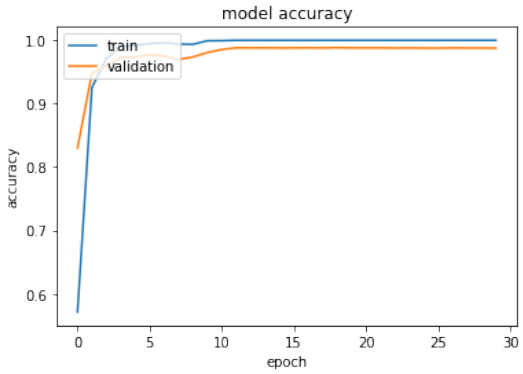


(b) Loss achieved during the training.

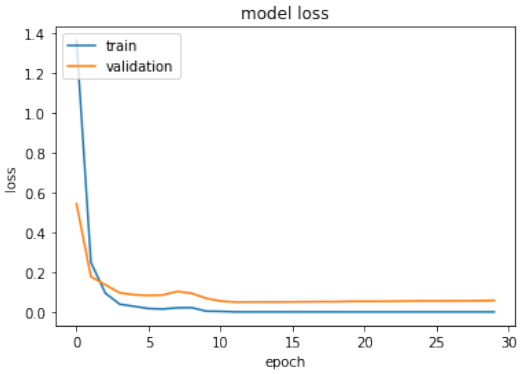
Figure 4: Training metrics behaviours.

3.3 Extended Training Set

In order to obtain good results also on the masked images, another training is performed. The training set is here extended to the union of the original training set with the 4 sets obtained by modifying the images belonging to the original training set as explained in section 2. Each of these sets contains less images w.r.t. the original set, because in those sets data augmentation is removed. The CNN and its parameters are kept unaltered, except for the epochs that are incremented to 30.



(a) Accuracy achieved during the second training.



(b) Loss achieved during the second training.

Figure 5: Training metrics behaviours on the second training.

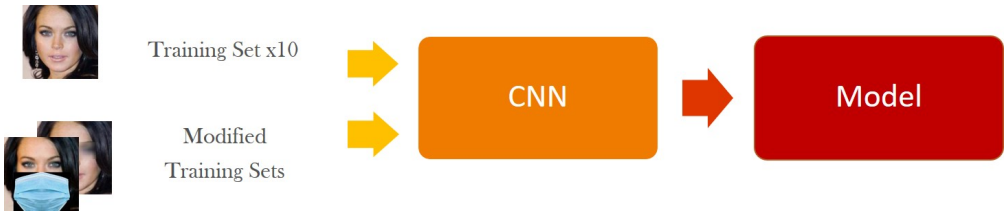


Figure 6: Extended model scheme.

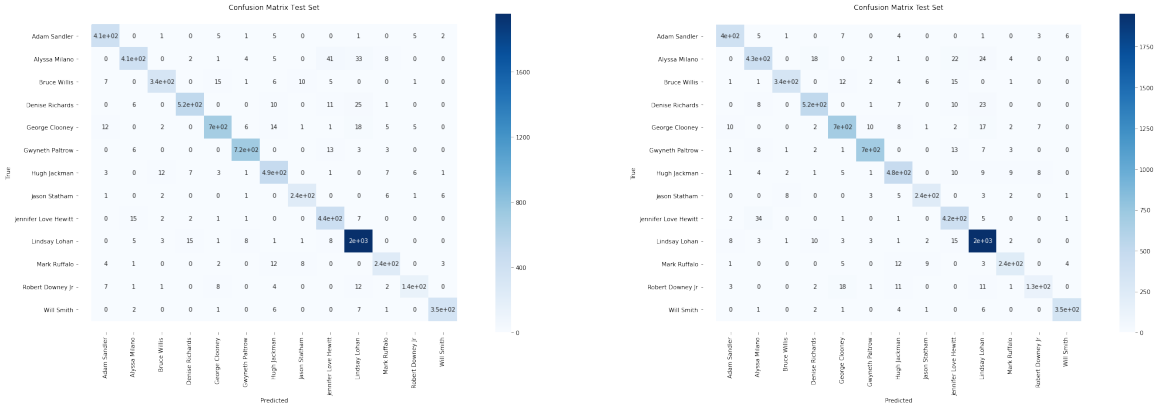
4 Results

The models are tested first on the original test set, then on the modified datasets and finally also on some videos of the involved people. The following paragraphs describe in detail the results obtained.

4.1 Images Test Set

The first set taken into account is the orginal test set of the faces, with no modifications. The set is fistly tested on the former model (section 3.2), achieving a value of accuracy of 93.15%, i.e. a pretty satisfactory result. Then, also the latter model (section 3.3) is tested on this set, achieving very similar results in terms of loss and accuracy (92.71%), as reported in table 1.

Figure 7 shows the confusion matrices resulting from the testings. On the rows are the true labels, while on the columns the predicted ones.
It is possible to notice that in the most of the cases people are well recognized, as the highes values are on the diagonal. In particular, *Lyndsay Lohan* achieves the highest values, but this is due to the fact that the set of her was the largest in terms of size.



(a) Base Training Set.

(b) Extended Training Set.

Figure 7: Confusion Matrices resulting from the test set.

	Base		Extended	
Test Set	**Accuracy**	**Loss**	**Accuracy**	**Loss**
Original	93.15%	0.3863	92.71%	0.5127
Blurred Eyes	63.76%	2.2732	89.68%	0.7752
Sun Glasses	50.21%	5.7826	88.53%	0.7614
Masked mouth	37.78%	8.1835	90.83%	0.7505
Coronavirus	28.82%	17.6496	86.51%	0.7636

Table 1: Table showing the metrics achieved by the CNN on the different test sets.

4.2 Modified Test Sets

The models are then tested on the modified test sets built in the pre-processing phase (section 2).

4.2.1 Base Model

The base model (section 3.2) is tested first, and the confusion matrices obtained testing the four sets are shown in figure 8. From the figure and from table 1 it is possible to notice that the results are very poor. In particular:

1. The *Glasses simulation* set, obtained blurring the region of the eyes, achieves an accuracy of 63.76%, which is the less worse among these four sets. This because the blurring has limited effects on the face, and the person is almost still recognizable. The confusion matrix (fig. 8a) is still almost diagonal, but there are also other elements with high values, e.g. *Lyndsay Lohan* is in many cases recognized as *Gwyneth Paltrow*.

2. The *Sunglasses simulation* set, where eyes are partially masked with black circles (*alpha* value < 1) achieves an accuracy of 50.21%, worse than the *Glasses* as the region covered is more influent. In the matrix (fig. 8b) has still the diagonal filled, but many people are often recognized as others (e.g. *Gwyneth Paltrow*), while some of them are more robust to this modification (e.g. *George Clooney* or *Lyndsay Lohan*).

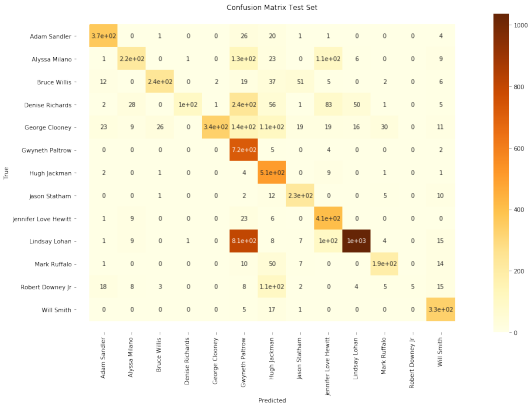
3. The *Mask simulation* set, where a black rectangle is on the mouth achieves an accuracy of 33.22% and made exception for *Lyndsay Lohan* or *Alyssa Milano* (fig. 8c), people are wrongly classified.

4. The *Coronavirus Medical Mask simulation* set, is finally a disaster, as it covers more than half of the face. It achieves an accuracy of 28.82% and the people are not recognized well.

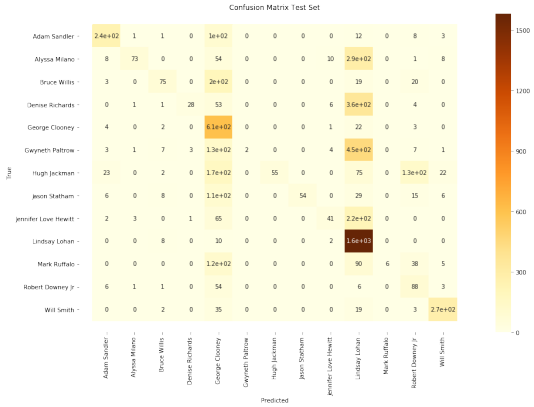
4.2.2 Extended Model

Using insead the extended Model, i.e. the one trained also with a subset of the modified samples (section 3.3), the results have a huge improvement on all the sets (table 1).
As a matter of fact, training the network with those samples[2] makes the model learn to recognize also the covered faces, and the confusion matrices in figure 9 prove the result.

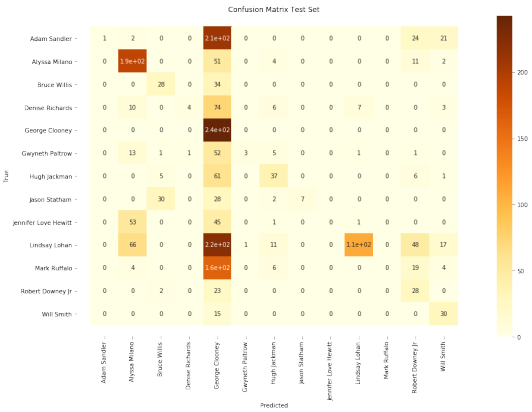
[2]These sets hold less samples because only the ones detectable with *Viola and Jones* after the modification are kept.



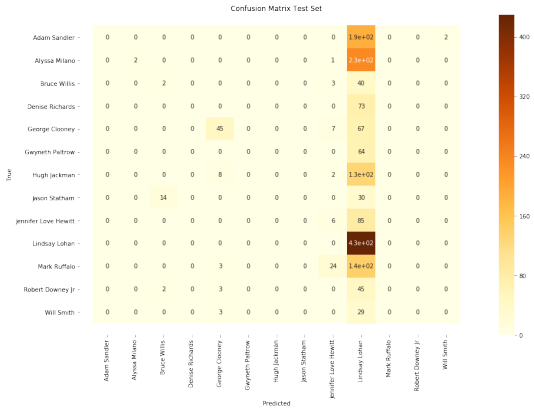
(a) Confusion Matrix of the Glasses simulation dataset



(b) Confusion Matrix of the Sunglasses simulation dataset

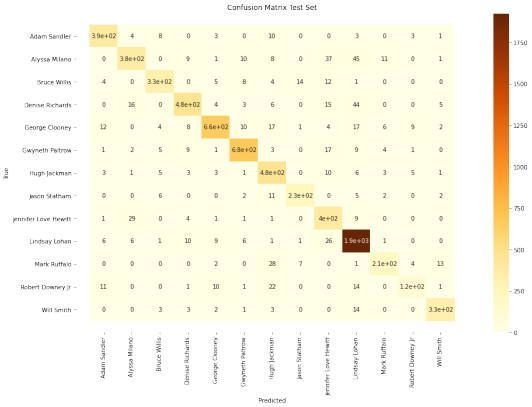


(c) Confusion Matrix of the Mask simulation dataset

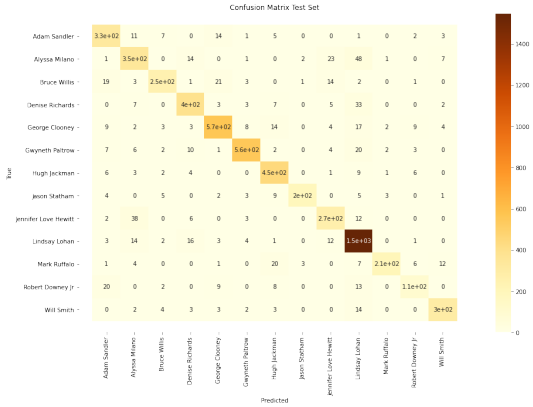


(d) Confusion Matrix of the Coronavirus Medical Mask simulation dataset

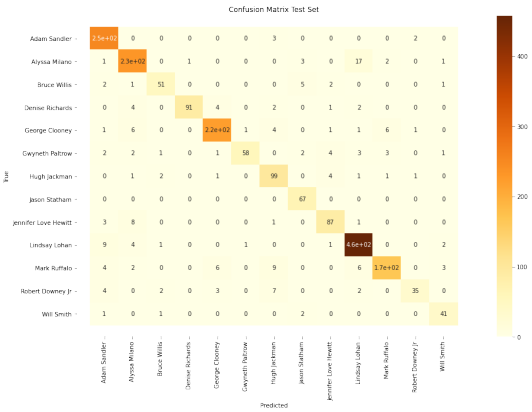
Figure 8: Confusion matrices of the modified datasets, Base Training Model.



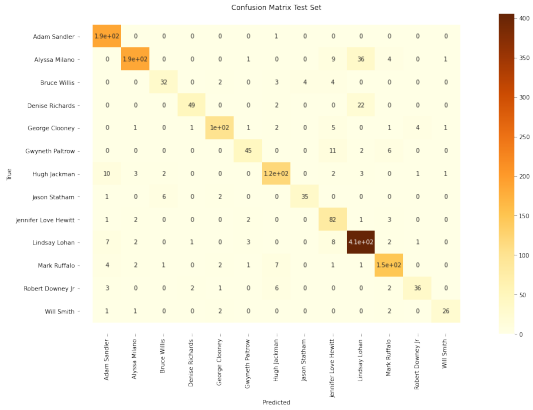
(a) Confusion Matrix of the Glasses simulation dataset



(b) Confusion Matrix of the Sunglasses simulation dataset



(c) Confusion Matrix of the Mask simulation dataset



(d) Confusion Matrix of the Coronavirus Medical Mask simulation dataset

Figure 9: Confusion matrices of the modified datasets, Extended Training Model.

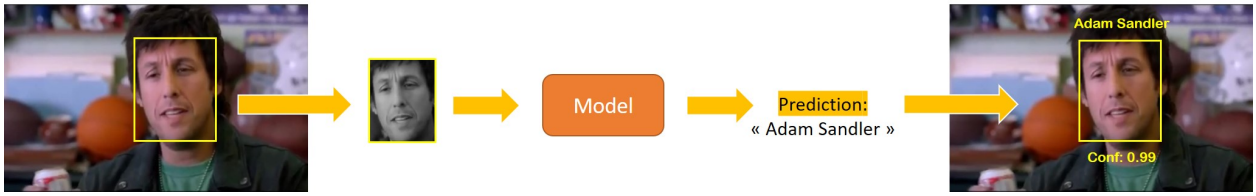


Figure 10: Video processing. In each frame, the faces are extracted and fed to the model, which makes the predictions.

4.3 Videos

The face recognition system starting from videos has been developed in **Python** using *OpenCV* and the video capture module within it. Figure 10 briefly visually summarizes the main passages it performs. In detail:

1. The frames are extracted sequentially from the video and computed one at a time. The frame rate of the videos is kept unaltered at $fps = 30$.

2. *Viola and Jones algorithm* is used to extract all the faces present in each frame, like what has been done for image pre-processing (section 2).
   This step allows the recognization of more than one person of the dataset, also if they are present in the same video frame.

3. The faces are normalized, as in the pre-processing; the model is fed with each of them and makes the predictions.

4. To skim the results the *confidence* is computed as:

   $$conf = max\{\mathbf{pred}\}$$

   where **pred** is the vector of the predictions.
   Then, only faces with $conf > 40\%$ are kept and the others are discarded.

5. Each frame is then added to an output video. If one or more faces are recognized, a rectangle is drawn around each of them, and if the face is accepted (i.e. it has $conf > 40\%$) a label is associated to it and printed close to the rectangle. An example result is shown in figure 11.

6. As the video is being processed, the predictions are stored in a list, ordered according to the frames temporal order, and keeping more than one label in the same cell if the frame holds more than one face.

4.3.1 Refinements

The predictions found this way are acceptable, but frequently can happen that people not present in the video are predicted. These predictions can be seen as *spikes of noise* in the predictions vector, i.e. predictions with high confidence, but predicted in a single frame and/or in very far frames in terms of time. In many situations the wrong prediction is due to the orientation or facial expression the person assumes.

In order to smooth these errors and give a final prediction on the whole video, some additional refinements are performed:

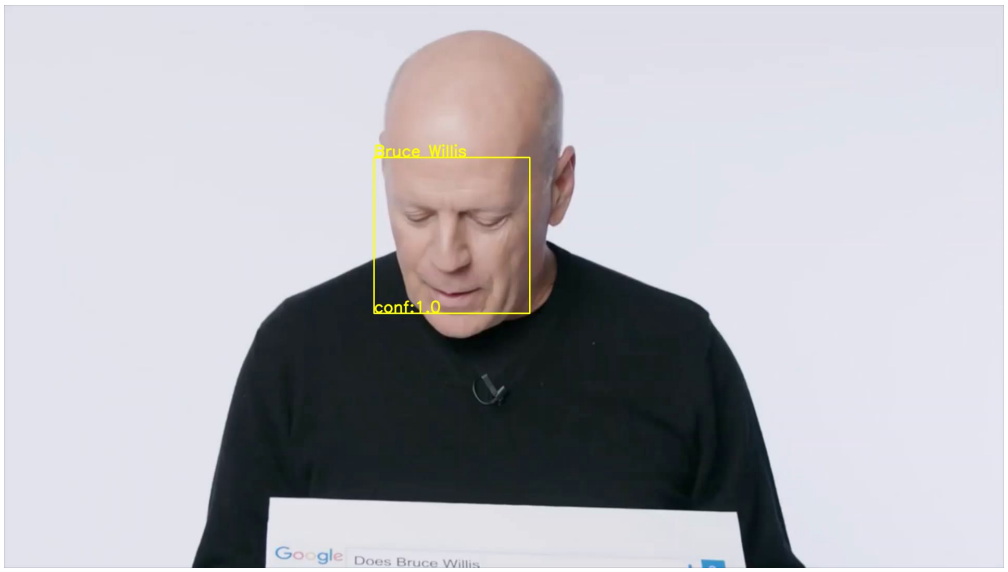


Figure 11: Frame from the output video, after detecting and computing faces.

- A threshold n_1 is set to limit the portion of the video in which people are present, i.e. only people that are predicted in more than $1/n_1$ of the total video length are kept. This trick works for short videos, almost of the same durations. Alternatively this threshold can be set to limit the number of the frames, with no relation with the video duration (e.g. the maximum number of predictions must be $n_1 = 30$, which corresponds to 1 second of video).
- The second threshold n_2 is the number of consecutive frames in which the person appears. This avoids the spikes of noise, as it is very difficult for a person to appear in the video only for 1 or 2 consecutive frames, i.e. for less than 1 second, in the video.

5 Conclusions

The face detection and recognition system built in this project is pretty simple, but achieves anyway satisfactory results. Some improvements could be obtained working more on the *pre-processing* step, e.g. normalizing the orientation and facial expression of people. In particular, detecting facial landmarks the best pre-processing results could be achieved.

From the results obtained we can state that a simple network trained on almost 30.000 images (The training set is composed of almost 3.000 but they were augmented of a factor of 10) is enough to obtain a satisfactory level of accuracy on the test set.

However, testing the partial hidden faces test sets on the network trained only on the original training set, leads to poor results. On the other hand, very relevant improvements can be seen if the testing is led with the model trained with the extended set of data. This because this model has learned to recognize also partial faces and consequently it gets satisfactory results also on the modified test sets.

Discussing about the images, the results obtained from the extended training are expected as the half faces achieve worse results w.r.t. complete faces, even worse as the face is more covered.
The algorithm has less problems with guys, maybe because the guys chosen in the set are very different one another, while the women have many traits in common (e.g. *Lindsay Lohan* and *Gwyneth Paltrow* were confused by the cnn in some occasions). On the other hand, the people with more data (e.g. *Lindsay Lohan*) achieve always a more accurate result, because a large amount of data gives more precision in the classification.

As regards videos, the results achieved are not perfect because it is a more challenging task. Anyway we can consider them satisfactory enough especially if the testing is performed on videos showing only the people involved in the dataset and in high definition. On the other hand *Viola and Jones* is not always accurate, because in many situations, objects that are not faces are recognized as faces, and this carries imprecisions in the final prediction. In addition the recognization is more accurate if people are almost always frontal and with a neutral expression.
The final passages which makes the generic prediction on the whole video, on the other hand, aims at limiting the wrong predictions and wisely tuning the parameters n_1 and n_2 , very good results are often achievable.

References

[1] Simone Milani, *Digital Forensics course slides*, 2020.

[2] Python tutorial on face recognition and identification
`https://www.youtube.com/watch?v=PmZ29Vta7Vc`

[3] Michele Rossi, *Lab Nootebooks, Human Data Analytics course*, 2020.

[4] Pietro Zanuttigh, *"Keras Tutorial", Computer Vision course*, 2020.

[5] OpenCV, *Viola and Jones algorithm* documentation
`https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_objdetect/py_face_detection/py_face_detection.html`